Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model

Year: 2022 | *Citations: 803* | *Authors: Shaden Smith, M. Patwary, Brandon Norick*

Abstract

Pretrained general-purpose language models can achieve state-of-the-art accuracies in various natural language processing domains by adapting to downstream tasks via zero-shot, few-shot and fine-tuning techniques. Because of their success, the size of these models has increased rapidly, requiring high-performance hardware, software, and algorithmic techniques to enable training such large models. As the result of a joint effort between Microsoft and NVIDIA, we present details on the training of the largest monolithic transformer based language model, Megatron-Turing NLG 530B (MT-NLG), with 530 billion parameters. In this paper, we first focus on the infrastructure as well as the 3D parallelism methodology used to train this model using DeepSpeed and Megatron. Next, we detail the training process, the design of our training corpus, and our data curation techniques, which we believe is a key ingredient to the success of the model. Finally, we discuss various evaluation results, as well as other interesting observations and new properties exhibited by MT-NLG. We demonstrate that MT-NLG achieves superior zero-, one-, and few-shot learning accuracies on several NLP benchmarks and establishes new state-of-the-art results. We believe that our contributions will help further the development of large-scale training infrastructures, large-scale language models, and natural language generations.